DOCUMENT RESUME

ED 457 235 TM 033 334

AUTHOR	Altman, Daniel R.
TITLE	Constructing Good Tests: A Summary of Basic Precepts and Principles Presented in Measurement Textbooks.
PUB DATE	2001-04-00
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Measurement Techniques; Objective Tests; *Rating Scales; Student Evaluation; *Test Construction; Textbooks
IDENTIFIERS	*Paper and Pencil Tests

ABSTRACT

Measurement is a necessary component for research in psychology and education. Researchers use various tests as tools to measure a given construct, making it necessary to have tests that assess a given construct accurately. There are specific principles necessary to follow when constructing a test. This paper discusses the principles for constructing two types of tests: paper and pencil tests and rating scales. The paper contains specific advice for constructing these types of paper and pencil tests: short-answer, true/false or alternative response, multiple-choice, and essay. The paper also contains advice about constructing numerical, graphic, and descriptive rating scales. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 457 235

Running Head: CONSTRUCTING GOOD TESTS

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

D.R. Altman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Constructing Good Tests: A Summary of Basic Precepts And Principles Presented in Measurement Textbooks

Daniel R. Altman

Texas A&M University 77843-4225

TM033334

Poster presented at the annual meeting of the American Educational Research Association, Seattle, April, 2001

BEST COPY AVAILABLE

Abstract

Measurement is a necessary component for research in psychology and education. Researchers use various tests as tools to measure a given construct. As Thompson (1994) states we are seeing "more and more regulatory and legislative mandates for the use of tests in making a variety of high-stakes decisions" (p. 837). Therefore, it is necessary to have tests that accurately assess a given construct. There are specific principles necessary to follow when constructing a test. This paper discusses the principles for constructing two kinds of tests: paper and pencil tests and rating scales.

Test Construction

Measurement is a necessary component for research in psychology and education. Researchers need a concrete method to describe, explain, predict, differentiate, diagnose, and make decisions. Measurement does this by gathering information and making that information meaningful with the powerful application of mathematics. The use of mathematics with measurement eliminates ambiguity, thus making the information meaningful, by creating an isomorphic relationship between the measured object and the numbers assigned to the measured object (Pedhazur & Schmelkin, 1991).

Researchers use various tests as tools to measure a given construct. As Thompson (1994) states we are seeing "more and more regulatory and legislative mandates for the use of tests in making a variety of high-stakes decisions" (p. 837). Therefore, it is vital that tests accurately assess information so that researchers can come to some meaningful conclusions. If tools that are flawed gather information, the information is useless. However, it is important to always remember that tests are not reliable, scores are (Thompson, 1994; Vacha-Haase, 1998; Thompson & Vacha-Haase, 2000).

It is absurd to say there is a <u>best</u> method for testing. Rather, there may be a <u>better</u> method for testing a given construct in a given situation. The goal of choosing which method to use is to consider which method might accurately measure a given construct under a given situation. Determining

this cannot be done effectively without the use of a theory to guide measurement (Pedhazur & Schmelkin, 1991).

Without theory guiding measurement, measurement haphazardly gathers important information at best. At worst, it gathers useless information wasting valuable time and resources while possibly influencing important decisions based on incorrect information. For instance, it is meaningless to measure personality without a theory about how personality develops or about various attributes belonging to personality. Therefore, measurement must be guided by a theory regarding the functioning or relations among a given set of variables that pertains to the overall construct being examined (Pedhazur & Schmelkin, 1991).

Purpose of the Test

Tests can be used for a variety of purposes, which must be determined before construction begins. The purpose of the test will depend on the theory guiding the construction of the test. Tests can be used to make placement, selection, instructional, evaluation, and treatment decisions in addition to reporting progress (Thorndike, Cunningham, Thorndike, & Hagen, 1991).

Tests can be used to assess an individual's knowledge, skills and abilities. This is useful when making decisions about placing individuals so that their knowledge, skills and abilities are used effectively. For instance, teachers often test students to determine whether or not a given student would be a more effective learner in a given class (e.g., remedial or advanced). Likewise, employers test workers to decide where they might be more productive. When using tests to make placement decisions be

sure the tests assess specific entry knowledge and skills for a given subject and remember that alternative placements differ, often dramatically (Gronlund, 1976; Thorndike et al., 1991). A child who is "misplaced" may be hindered dramatically academically and socially; workers who are placed in an area beyond or beneath their skills level may become unproductive.

Similarly, tests can be used to determine whether a person should be admitted into a given organization. Thorndike et al. (1991) states, "The purpose of a system for selection decisions is to hire or admit those individuals who will maximize the total value to the institution." A university that wants the prestige of being highly rated must admit those students whose G.P.A. and G.R.E. or S.A.T. scores are in the higher end of percentile rankings.

In addition, tests can be used to determine knowledge deficits (Thorndike et al., 1991). The information gained is used to guide the treatment (e.g., classroom curriculum). This is important in a classroom setting where a teacher needs to plan a lesson according to how much the students know. However, this task becomes difficult when there are no defined boundaries in a given domain (Thorndike et al., 1991). Therefore, when constructing a test consider what is important for the test-taker to know or not know.

Tests can be used to determine the effectiveness of a particular treatment (e.g., therapy, instruction, etc.) or of an individual's skills and abilities (e.g., performance evaluation) (Thorndike et al., 1991). This kind of testing is sometimes

referred to as formative testing (Gronlund, 1976). The decisions based on the results of formative tests can produce change that benefits both the individual and the organization.

Tests also can be used to tailor a given treatment to the test-taker. For instance, various psychological tests (e.g., Beck's Depression Inventory, the M.M.P.I., etc.) help illumine various characteristics of a psychological problems. This kind of a test is called diagnostic testing and includes many items (Gronlund, 1976).

Finally, tests can be summative, determining outcome of a given treatment (e.g., therapy, instruction, etc.). This kind of testing is often called summative testing (Gronlund, 1976). Often a given treatment or program receives funding as long as those to whom the treatment is applied benefit. For instance, an insurance company will withhold reimbursements if there is no evidence of growth from those receiving treatment. The results of summative tests require a standard, which will depend on the importance of the outcome (Thorndike et al., 1991).

Therefore, before constructing a test determine its purpose. Once the purpose has been identified one can operationally define the construct being examined. Make the construct as specific as possible in order to stay focused (Pedhazur & Schmelkin, 1991).

One way to facilitate the definition for the construct is to develop a table of specifications (Gronlund, 1976). A table of specifications is a simple way to outline the objectives for the test-taker regarding construct being measured (see Table 1). Notice this table is a two-way chart that relates the objectives

to the content area. This chart also specifies the nature of the desired test sample (Gronlund, 1976).

Each objective must be specific and measurable. Objectives may be categorized into various domains: cognitive, affective, and psychomotor (Sax, 1989). See Table 2 for an example of the cognitive domain.

Once the test content has been determined the kind of test can be selected. There are two kinds of tests that will be discussed in the remainder of this paper: paper and pencil and rating scales. Each kind of test has advantages, disadvantages and specific rules.

Paper and Pencil

The paper and pencil test is the most commonly used test. The test-taker generally does not rely on someone to give instructions for each item, making this test the easiest to administer. If test-takers have adequate reading skills, they can read each item and follow the instructions. There are four kinds of test items commonly used in the paper and pencil test: short-answer, true/false or alternative response, multiple-choice, and essay

The short-answer item is considered a simple form test item (Gronlund, 1976). This item is used to measure relatively simple learning objectives (e.g., basic knowledge). This item can easily be used to assess knowledge of terminology, specific facts, principles, methods, or interpretations of data (Gronlund, 1976).

Short-answer items are relatively easy to construct and to evaluate. A great advantage of these items is that the test-taker must supply the answer. Therefore, there is no doubt that the test-taker has the knowledge. However, these items are limited to simple learning outcomes. Therefore, higher cognitive domains are neglected (e.g., analysis, synthesis, or evaluation). The evaluator must also decide whether to assess spelling (Gronlund, 1976).

The following suggestions facilitate writing short-answer items:

1. Word the item so that a brief and definite response is required (Gronlund, 1976; Sax, 1989). In this way, correct answers will be easy to distinguish from incorrect answers.

2. Limit the number of blanks to one (Gronlund, 1976; Sax, 1989). Too many blanks can be confusing for the test-taker.

3. Place the blanks for each item to the right of a question, making sure they are all of equal length (Gronlund, 1976; Sax, 1989). Not only does this facilitate scoring, but it also makes it easier for the test-takers. Blanks of unequal length may provide a clue to the answer of a question.

4. Do not use statements taken directly from the textbook (Gronlund, 1976; Sax, 1989). Items taken directly from the textbook encourages memory rather than real learning.

5. Use direct questions rather than incomplete statements (Gronlund, 1976). Direct questions provide structure and decrease ambiguity.

6. Avoid using specific determiner (e.g., a, an) as they may provide clues to the answer (Sax, 1989).

Alternative response items are used to identify the correctness of statements (Gronlund, 1976). A commonly used alternative response item is the true/false item.

Like short-answer items, alternative response items measure simple learning outcomes and are easy to construct. A wide sampling of material can be obtained with alternative response items. However, complex learning outcomes are neglected. Alternative response items are dependent on absolute judgments that may be unfair given that most "facts" are neither entirely true nor false. These items are also susceptible to guessing (Gronlund, 1976; Sax, 1989).

The following are suggestions for writing alternative response items:

1. Use direct, brief statements that are either entirely true or entirely false (Gronlund, 1976; Sax, 1989). Statements that are not entirely true or false cannot be answered adequately. Long, complex, general statements are confusing for test-takers.

2. Use positive statements (Gronlund, 1976; Sax, 1989). Negative statements, especially double negative statements, are confusing to read and to answer. The object is to find out what the test-takers know or do not know, not to confuse them.

3. Avoid introducing more than one idea in an item, unless a cause and effect relationship is being measured (Gronlund, 1976). Items with two ideas may be impossible to

answer correctly. One idea may be true while the other idea may be false.

4. Make sure there are an equal number of positive or negative responses (Gronlund, 1976; Sax, 1989). True items are easier to construct. Therefore, most alternative response tests contain more true items than false items.

Multiple-choice items are similar to alternative response items, only with more than two choices. Multiple-choice items consist of two parts: the stem and a number of alternatives. The stem is a statement or question that is answered by choosing the correct alternative. Alternatives that are incorrect or least appropriate are called distracters (Gronlund, 1976; Sax, 1989). Test-takers must decide which alternative is either correct or the best answer.

Multiple-choice items have a number of advantages. Multiple-choice items can be used to measure a variety of simple or complex learning outcomes. They can measure a large sample of material in a relatively short period of time since little writing is required of test-takers. Results of these items are amenable to diagnosis as examining wrong answers gives a greater idea of the test-taker's knowledge. It is relatively easy to construct high quality items. These items also reduce the effects of guessing (Gronlund, 1976; Sax, 1989). However, multiple-choice items are not without their limitations.

Multiple-choice items are limited at the verbal level. Test-takers with lower verbal ability might be disadvantaged. These items are not well suited for measuring some problem-

solving skills in science and mathematics. It is also difficult to assess the ability to organize and present ideas with these items (Gronlund, 1976). Finally, it is difficult to supply incorrect, but plausible distracters (Gronlund, 1976; Sax, 1989).

Multiple-choice items can either contain a stem that is a direct question or that is an incomplete statement. For example,

Direct question:

"Which paper and pencil item contains a number of alternatives:"

Incomplete statement:

"Multiple-choice items contain a stem and:"

Multiple-choice items can ask for an answer that is entirely correct, or that is the best answer (Gronlund, 1976). However, unless otherwise stated, test-takers should assume there is a defensible correct option (Sax, 1989).

The following are suggestions to facility writing multiple-choice items:

1. The stem should make sense and present a definite problem (Gronlund, 1976; Sax, 1989). It is difficult for test-takers to know what is expected from one or two word stems.

2. Include as much of the item in the stem as possible and omit irrelevant material (Gronlund, 1976). Irrelevant material confuses test-takers unnecessarily.

3. Use positive statements as much as possible (Gronlund, 1976; Sax, 1989). Use negative statements only when significant learning outcomes require them. Also, when using negative

statements, emphasize the negative word with all caps, underlining or italicizing. For example,

"Use positive statements in the stem EXCEPT:"

4. Make sure all the alternatives are grammatically consistent with the stem (Gronlund, 1976; Sax, 1989). Alternatives that are not grammatically consistent with the stem may provide the test-taker with clues.

5. While all alternatives should be plausible, only one should be a defensible correct or best option (Gronlund, 1976; Sax, 1989). Alternatives that are not plausible are easy to detect. However, if there are more than one defensible correct options, scoring will be problematic.

6. Verbal associations between the stem and the correct option should be avoided. Therefore, avoid specific determiners (Gronlund, 1976; Sax, 1989).

7. Correct options should randomly appear in each of the alternative positions (Gronlund, 1976; Sax, 1989). Test-takers unsure of a correct answer tend to select middle options.

8. Use options such as "none of the above" sparingly, while avoid using options such as "all of the above" altogether (Gronlund, 1976; Sax, 1989). "All of the above" can be ruled out easily for alternatives requiring a best answer.

9. Avoid measuring opinions because they are neither correct nor best (Sax, 1989).

10. Avoid using alternatives that overlap in meaning (Sax, 1989).

Like short-answer items, essay items require test-takers to supply the answer to a given question. However, while short-answer items require a brief response and measure simple learning objectives, essay items require a longer response and measure simple and complex learning objectives (Gronlund, 1976; Sax, 1989). Essay items are relatively easy to construct and emphasize the integration and application of thinking and problem-solving skills. These items give students freedom to respond within broad limits (Sax, 1989). It is also difficult for test-takers to guess at these items.

However, essay items are difficult to score objectively (Gronlund, 1976; Sax, 1989). Test-takers who have poorer handwriting may receive lower scores. Essay questions are also time-consuming for both the evaluator and test-taker. Because test-takers have the freedom to respond they may not discuss a given construct as detailed as the evaluator requires. Finally, essay items reduce guessing but do not prevent bluffing (Gronlund, 1976; Sax, 1989).

The following are suggestions to facilitate writing essay items:

1. Specify limitations for each response (Gronlund, 1976; Sax, 1989). Test-takers should be told how much time or space to allot each response along with the number of points assigned to each.

2. Ask direct questions to avoid ambiguity. Questions should reflect a given learning objective (Gronlund, 1976; Sax, 1989).

3. Make each item relatively short and increase the number of items in order to facilitate a thorough assessment of the test-taker's knowledge and skills (Sax, 1989).

4. Avoid the use of optional questions as much as possible (Gronlund, 1976; Sax, 1989). Optional questions do not allow the evaluator to assess the test-takers knowledge and skills as the same level.

Rating Scales

Ratings are observations judged according to given criteria. These are usually recorded with the use of a scale that breaks down the criteria into various components (Gronlund, 1976; Pedhazur, & Schmelkin, 1991; Sax, 1989). Rating scales fall into two categories: process scales and product scales (Sax, 1989). Process scales are used to judge procedures in performing a task, while product scales are used to judge results of performing a task (Sax, 1989). Various scales exist under each of these two categories.

Rating scales can either be unstructured or structured (Sax, 1989). Unstructured scales lack criteria to guide the rater. For instance, a letter of recommendation is an unstructured rating. These are usually unsatisfactory unless the rater includes ratings from opposite poles (e.g., strengths and weaknesses) (Sax, 1989). Structured scales include specific criteria used to guide the rater.

Most ratings are structured scales. These scales usually have a characteristic that is measured by a high to low scale. In other words, the high end of the scale describes one portion

of the characteristic, while the low end of the scale offers the opposite. For instance, a scale measuring temperature might have "hot" at one end and "cold" at the other end. A scale for a given behavior might have "always" at one end and "never" at the other. A scale measuring attitudes might have "strongly agree" at one end and "strongly disagree" at the other end. The rater, uses any point in between the two ends of the scale to describe that which is being rated.

There are three kinds of rating scales that are commonly used: numerical, graphic and descriptive rating scales (Gronlund, 1976; Pedhazur, & Schmelkin, 1991; Sax, 1989). Each of these scales has common characteristics. They each are based on the perception of the rater. Because of this, rating scales are vulnerable to misinterpretation (Pedhazur, & Schmelkin, 1991). Each scale has some sort of criteria describing the categories. Unfortunately, raters often are not given the criteria for the categories, which leads to inaccurate ratings (Pedhazur, & Schmelkin, 1991). Finally, rating scales often have a composite score that combines all the scores from the rating scale (Pedhazur, & Schmelkin, 1991).

Numerical rating scales assign a number to a descriptively ranked category (Sax, 1989). One of the most popular numerical rating scales is referred to as a Likert-type scale (Pedhazur, & Schmelkin, 1991). When using a numerical rating scale it is always important to define each rating. For instance, "1" might represent "extremely dissatisfied," "2" might represent "somewhat

dissatisfied," all the way to "7" which might represent "extremely satisfied."

When summing these responses it is important that opposite words (e.g., negative verses positive) are assigned the same value. For instance, if "extremely satisfied" is given the value of 6, "extremely dissatisfied" must also be given a value of 6 (Pedhazur, & Schmelkin, 1991).

Graphic rating scales consist of either a vertical or horizontal line used to measure a given characteristic (Gronlund, 1976; Sax, 1989). The rater places a mark somewhere on the line that closely describes that which is being rated. Ratings may be more accurate if there is an example at various points on a vertical line the rater can use as a guide (Sax, 1989). The graph of the Global Assessment of Functioning (GAF) Scale commonly found in the DSM is an example of this kind of scale.

Finally, descriptive rating scales consist of descriptive phrases or words (Gronlund, 1976; Pedhazur, & Schmelkin, 1991). One example of this is the Semantic Differential (SD) technique. The idea behind the SD technique is that concepts evoke a variety of meanings that can be categorized into a few dimensions (Pedhazur, & Schmelkin, 1991).

These dimensions are: evaluative, potency, and activity (Pedhazur, & Schmelkin, 1991). Words used for the evaluative domain might be "good" or "bad," "nice" or "mean," "fair" or "unfair." Words describing potency might be "weak" or "strong," "soft" or "hard." Words describing activity might be "fast" or "slow," "hopped" or "leaped."

When constructing a scale such as the SD technique, there may be a tendency to place words of a given pole (e.g., negative) in the same column. Therefore, in order to minimize response set, it is important to randomly alternate these words (e.g., place "nice" and "unfair" in the same column) (Pedhazur, & Schmelkin, 1991).

Rating scales are often used not only to gain information from observations, but also to guide observations. When observing a given behavior or product, rating scales help observers know exactly what they are observing (Pedhazur, & Schmelkin, 1991).

Summary

Measurement is a necessary part of life. In research, tests are commonly used to gain information necessary for measurement. As this paper has indicated, test construction is sometimes a long, arduous process. However, it is also a necessary process. In order to gain information accurately, time and planning must be a part of the process. Without effective planning the information gained from tests is useless, yielding analysis of inaccurate information which can lead to incorrect conclusions.

References

Gronlund, N.E. (1976). Measurement and evalutation in teaching (3rd ed.), NY: Macmillan.

Pedhazur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sax, G. (1989). Principles of educational and psychological measurement and evaluation (3rd ed.). Belmont, CA: Wadsworth.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. Educational and Psychological Measurement, 60, 174-195.

Thorndike, R.M., Cunningham, G.K., Thorndike, R.L., & Hagen, E.P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: MacMillan.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Table 1

Table of Specifications for a Summative Third-Grade Social Studies Test

	Objectives					
Content	Knows Common Terms	Knows Facts	Understands Principles and Generaliza-tions	Applies Principles and Generaliza-tions	Interprets Charts and Graphs	Total
Food	3	5	2			10
Clothing	2	5	3			10
Transport a-tion	5	5	3	3	5	20
Communica -tions	5	2	3	2	5	17
Shelter	3		2	4		9
City Life	4	2	3	7		16
Rural Life	4	2	6	6		18
Total	26	20	22	22	20	100

Adapted from Gronlund (1976) page 141.

Table 2

Bloom's Taxonomy of Teaching Objectives: Cognitive Domain

Level	Instructional Format	Implicit Format
1. Knowledge		
of terminology	Writes meaning; selects definition	Knows/memorizes meaning
of facts	Reports on the…	Remembers…
of trends	Lists the sequence	Is aware of the sequence
of criteria	Recites criteria	Has knowledge of criteria
2. Comprehension		
Translation	Translates writing	Comprehends words
Interpretation	Lists implications	Interprets
Extrapolation	Predicts the outcome	Determines the outcome
3. Application	Solves problems	Applies knowledge
4. Analysis		
of elements	Circles examples of…	Distinguishes between…
of relationships	Lists fallacies in…	Can find fallacies in…
of principles	Infers relationship	Clarifies structure of…
5. Synthesis		
Production	Composes or proposes	Integrates or develops
Derives relations	Composes novel situation to…	Derives a set of relations to…
6. Evaluation (Judgments in terms of…)		
internal evidence	Makes judgments about…	Makes judgments about…
external criteria	Orally composes arguments against…	Composes arguments against…

Adapted from Sax (1989) pages 72-73.